Authorship Style Transfer with Inverse Transfer Data Augmentation

Anonymous ACL submission

Abstract

Authorship style transfer aims to modify the style of neutral text to match the unique speaking or writing style of a particular individual. While Large Language Models (LLMs) present promising solutions, their effectiveness is limited by the small number of in-context learning demonstrations, particularly for authorship styles not frequently seen during pre-training. In response, this paper proposes an inverse transfer data augmentation (ITDA) method, leveraging LLMs to create (neutral text, stylized text) pairs. This method involves removing the existing styles from stylized texts, a process made more feasible due to the prevalence of neutral texts in pre-training. We use this augmented dataset to train a compact model that is efficient for deployment and adept at replicating the targeted style. Our experimental results, conducted across four datasets with distinct authorship styles, establish the effectiveness of ITDA over traditional style transfer methods and forward transfer using LLMs. For further research and application, our dataset and code are openly accessible at https://github.com/AnonymousRole/Lifelike-Writer.

1 Introduction

Text style transfer, a technique that rewrites text into a specific style while retaining content, has gained attention in recent years. Most existing methods can only effectively address style attribute transfer, which shifts text on a particular style dimension, such as sentiment, formality and politeness. We refer to aforementioned style with well-defined attributes as polar style. Unlike these, authorship style (Xu et al., 2012; Carlson et al., 2018) is a unique category that describes an individual's writing or speaking style. It is characterized by word choice, structure, quirks, and topics but lacks well-defined attributes, making it difficult to categorize as positive/negative or polite/impolite. Figure 1 (a) displays some examples which clarify that authorship style involves more intricate and indefinable elements compared to polar style.

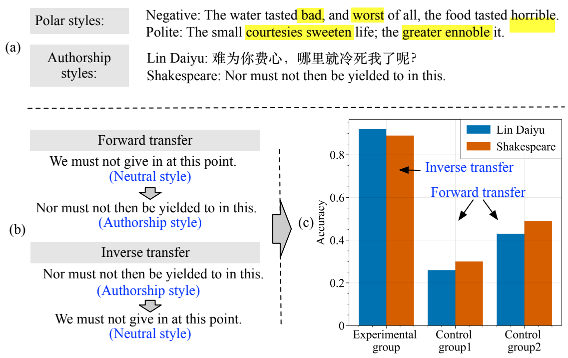

Figure 1: Illustration of (a) polar style and authorship style with well-defined stylized words highlighted; (b) forward transfer and inverse transfer; (c) experimental results of pilot study.

This paper investigates authorship style transfer, which aims to transform neutral style text into text matching a specific author's style, a topic previously addressed in studies like (Syed et al., 2020) and (Patel et al., 2022). This problem offers diverse applications, including creating personalized digital assistants that communicate in a user's chosen style, aiding students and researchers in understanding different authors' unique writing styles—important for literary studies and education—and improving privacy by altering an individual's writing style to conceal their identity, particularly useful for sensitive documents.

Recently, Large Language Models (LLMs) such as GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) have been utilized for their strong generalization abilities to infuse desired styles into generic neutral texts—a process known as **forward transfer**—through in-context learning with a few demonstrations. The limited input length of LLMs restricts the number of demonstrations possible, hindering comprehensive instruction on a target author's style, particularly for authorship styles not extensively covered in LLM pre-training. Research such as (Reif et al., 2022; Patel et al., 2022) suggests incorporating descriptive adjectives into prompts to capture the style of target author. While this method lessens the need for many demonstrations, distilling an author's unique style into just a few words remains a difficult task.

Instead of relying on in-context learning with limited examples to guide LLMs in authorship style transfer, we propose an alternative method: training a smaller, specialized model using abundant examples augmented from existing stylized texts. This method is more effective for dealing with uncommon authorship styles and also cuts down on costs related to model deployment and inference. The crucial part of this approach involves creating high-quality pairs of neutral and stylized text for training our compact model. Since it's possible to get text samples in the style of the target author, we've developed a method called **Inverse Transfer** Data Augmentation (ITDA). This method uses LLMs to remove the specific style from texts, turning them into neutral texts. These transformed texts are then used in reverse – from neutral to stylized – to train our compact model. This method of "inverse" data augmentation often works better than the usual "forward" approach, as LLMs are typically better at creating neutral rather than highly stylized texts due to the prevalence of neutral texts in pre-training. We illustrate this concept using diagrams in Figure 1 (b) and have conducted a pilot study, detailed in Section 4, showing the effectiveness of this inverse approach. The results, displayed in Figure 1 (c), show an impressive 40-66% increase in accuracy[1].

In implementing ITDA, our focus includes dynamic prompting and stylized text augmentation. Dynamic prompting is designed to identify the most appropriate prompts for each piece of stylized text, effectively aiding in the style removal process. This is achieved by clustering the corpus and assigning the most representative demonstrations to each cluster, enabling the selection of the most fitting prompts for stylized texts. Additionally, to tackle the challenge of limited availability of stylized texts in less common styles, we utilize LLMs to generate new texts in these specific styles. The key contributions of this paper are summarized as follows:

[1]We measure accuracy with a style classifier, and more information can be found in Section 4.

- We propose ITDA, an inverse transfer data augmentation method designed to address authorship style transfer. Leveraging LLMs, we perform inverse transfer to convert stylized texts into neutral texts, resulting in a corpus that trains a compact and deployable model.

- We introduce a clustering-based dynamic prompt selection method to bolster the performance of inverse transfer. We also leverage LLMs to synthesize new texts in the target style to mitigate data scarcity.

- Through comprehensive experiments conducted on four authorship-stylized datasets in both Chinese and English, we demonstrate the advantages of ITDA compared to traditional style transfer approaches and forward transfer on LLMs.

2 Related Work

Style transfer methods can be roughly classified into three categories: original representation revision, latent representation revision, and in-context learning on LLMs. The first two methods are predominantly utilized for style attribute transfer, with several works also applying to authorship style transfer.

Original representation revision (Sudhakar et al., 2019; Reid and Zhong, 2021) follows a "delete-generate" framework (Li et al., 2018), in which the original stylized words are removed and the desired stylized words are added. While offering excellent interpretability by modifying original words, this approach struggles with authorship style transfer, as identifying stylized words within the authorship-style text is challenging.

Latent representation revision (Wang et al., 2019; Xu et al., 2020; Xiao et al., 2021) involves revising the original text's latent representation within a Euclidean space, guided by content and style loss, and then decoding to generate the target-stylized text. (Syed et al., 2020; Riley et al., 2021) explore its application in authorship style transfer. However, directly manipulating the latent representation may lead to a low-density region, resulting in unpredictable and low-quality text output. Besides, this method of revising the latent representation lacks fine-grained control over the target style (Jin et al., 2022).

In-context learning using LLMs is currently a favored method for style transfer. A prime example is the Prompt-and-Rerank technique with GPT-2 (Suzgun et al., 2022), which generates multiple outputs for each input and ranks them based on factors like textual similarity, style, and fluency. Researches like (Patel et al., 2022) and (Reif et al., 2022) incorporate descriptive adjectives extracted from stylized texts into prompts in GPT-3.5 to mimic a target author's style. The former applies the same demonstrations across different styles, while the latter varies them according to the style. However, distilling an author's style into a few words is complex, and the limited demonstrations may not fully capture the nuances of less common styles. While the latter also uses inverse transfer, their focus is on automating demonstrations rather than data augmentation to provide a compact model with more extensive training examples.

3 Problem Definition

Authorship Style. Neutral text involves writing that is devoid of a particular style of interest. Neutral text is prevalent across various types of articles and platforms in reality. This is exactly why we select it as the transfer target. Stylized text, on the other hand, contains distinctive expressive elements, such as sentiment and formality. Authorship style is a special type of stylized text which embodies an individual author's unique word choices, writing structures and emotional inclinations. However, unlike other well-defined styles, the authorship style lacks clearly defined attributes, making it challenging to summarize its characteristics in a few words.

Authorship Style Transfer. Given a target authorship style s, and an input text x with the neutral style, our objective is to transform it into text y that exhibits the style s. We refer to this conversion process as **forward transfer**. Conversely, the process of converting y back to x, where the style s is removed from y, is termed **inverse transfer**. We use the notation D^s to represent a collection of texts that exhibit an authorship style s.

4 Pilot Study

As analyzed in Section 1, LLMs are more skilled at inverse transfer rather than forward transfer. We design the following controlled experiments to validate this assumption.

Datasets. We prepare two distinct authorship-stylized datasets. The first style embodies the essence of "Lin Daiyu", an iconic figure from Chinese ancient literature, while the latter style captures the essence of "Shakespeare", a renowned English playwriter. The two datasets consist of 1,000 and 4,000 textual pieces respectively.

Experimental Protocol. We devise the experimental group for inverse transfer and the control group for forward transfer, employing the few-shot prompting technique on GPT-3.5 to validate our hypothesis. For both author-stylized datasets, we repectively select 8 sentences, denoted by $\{y\}$, and manually transcribe their corresponding neutral text $\{x\}$. These are paired to form $\{(y, x)\}$, which serves as the demonstrations for inverse transfer. Then we inverse them to form $\{(x, y)\}$, which are used as the demonstrations for forward transfer.

In the experimental group, the input stylized text is collected from the remaining sentences of stylized datasets, excluding those chosen as demonstrations. In the control group, the input neutral text contains two types. The first involves random topics and the second involves similar topics with the authorship-stylized dataset. We choose two control groups because we have observed a correlation between the performance of the forward transfer and the topics of the input neutral text. If the topics significantly diverges from author-stylized dataset, the forward transfer process becomes challenging. To ensure a fair comparison between the experimental and control groups, we strive to align the topics of the inputs to the forward transfer with authorship-stylized datasets as closely as possible. Details regarding the construction of neutral texts can be found in Subsection 6.1.

Observation. We measure inverse and forward transfer accuracy by pre-trained binary classifiers tailored to identify the given authorship style "Lin Daiyu" and "Shakespeare". The training setup can been seen in Subsection 6.1. The accuracy of an output of the inverse transfer is assigned a value of 1 if its classification result is negative, and 0 otherwise. Similarly, the accuracy of an output of the forward transfer is marked as 1 if its classification result is positive, and 0 otherwise.

Figure 1 (c) illustrates that, in comparison with the experimental group for inverse transfer, both control groups for forward transfer underperform by 40-66% accuracy. We conjecture that neutral text, with its simpler form, is relatively easy to learn. During pre-training, LLMs are exposed to a greater volume of neutral text than specific au-

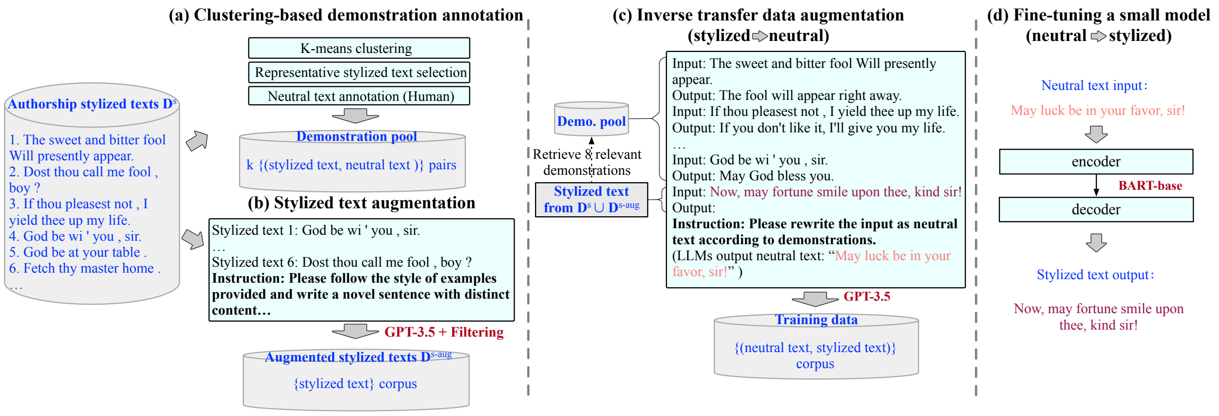

Figure 2: The ITDAframework, featuring three key components: (a) Clustering of D^s with annotation of representative texts for dynamic prompting and (b) augmentation of stylized texts to create D^{s-aug}. (c) Inverse transfer of stylized texts $D^s \cup D^{s-aug}$ to neutral texts using dynamic prompting by LLMs to create augmented parallel data. (d) Fine-tuning a compact model with the augmented parallel data.

thorship style text. This increased exposure augments the ability of LLMs to generate neutral text. Guided by this observation, we craft our inverse knowledge distillation method for authorship style transfer. This observation offers crucial supporting evidence for the inverse transfer data augmentation method that we propose subsequently.

5 ITDA

5.1 Framework Overview

The basic idea of ITDA is to augment data by inverse transfer on LLMs and then fine-tune a small model based on this augmented pairs. The framework comprises three essential components, as illustrated in Figure 2. Note we apply the framework to train a separate compact model for each style s.

This framework surpasses the direct few-shot prompting for forward transfer, primarily due to the input length constraints of LLMs. Given the intricate nature of authorship style, effectively transferring arbitrary neutral text demands a sufficient number of $\{(x, y)\}$ pairs to facilitate a comprehensive understanding of the authorship style by LLMs. Unfortunately, the length limitation prevents the inclusion of a large number of examples, potentially prompting LLMs to draw style inferences from their pre-existing knowledge beyond the limited demonstrations. For instance, if the target is to transfer text into style of "Lin Daiyu", LLMs may inadvertently mirror a classical Chinese style rather than the specific style of "Lin Daiyu". Similarly, when aiming to emulate a "Shakespeare" style, LLMs may unintentionally reflect an archaic English style. Unlike the direct forward transfer, we opt for the easier inverse transfer process (c) to create $\{(x, y)\}$ pairs and train a compact model (d) to enable exposure to a greater amount of training examples.

Besides (c) and (d), we further propose two enhancement strategies (a) and (b). The first strategy involves replacing the original static prompts with dynamic prompts to improve the conversion of given authorship-stylized text into neutral text. This reduces the likelihood that the LLMs will infer based on their pre-existing knowledge. More specifically, we adopt a clustering-based method to match optimal prompts for each input stylized text. The second strategy focuses on data augmentation for the input authorship-stylized text. Since the collected authorship-stylized text is often limited, we leverage LLMs to synthesize additional authorship-stylized text, thereby enhancing the model's ability to handle diverse scenarios.

5.2 Clustering-based Demonstration Annotation

To enhance the capability of LLMs in inverse transferring text with varied authorship styles, we construct a demonstration pool to dynamically assign prompts for each piece of authorship-stylized input. Optimal demonstrations are those that mirror the input's key attributes like phrasing, sentence structure, and rhetorical elements, contributing to a coherent language style match.

To select representative demonstrations and reduce human labor, we introduce a clustering-based strategy. Although this pool is much smaller than D^s, it's carefully designed to encapsulate the given authorship style, thus offering an effective solution.

The clustering-based prompting technique that we adopt is validated by (Zhang et al., 2022; Li et al., 2023), confirming that the chosen demonstrations from different clusters are diverse enough to facilitate the inference of a wide range of new input.

Typically, we construct the demonstration pool for style s in the following manner: (1) We first use Sentence-BERT (Reimers and Gurevych, 2019) to represent each sentence $y \in D^s$, then apply the k-means algorithm to cluster them into k categories. Calculation details for k can refer to Appendix B; (2) Then, we select the center of each cluster as a representative text and pair it with its counterpart in neutral style to form the demonstration pool. The counterpart is first generated by LLMs and then refined by humans;

5.3 Stylized Text Augmentation

Collecting adequate text in a specific authorship style can be challenging, especially when datasets that align with such styles are scarce or unavailable as open-source datasets. To overcome this limitation, we leverage LLMs to generate new texts in accordance with the target authorship style as D^{s-aug}, yet encompass distinct content. We take 6 sentences from D^s and combine them with the instruction such as "*Please follow the style of examples provided and write a novel sentence with distinct content. The newly generated text needs to cover a wide range of topics across various fields.*" This serves as a prompt to guide the LLM in generating new text. Different texts from D^s can be substituted as prompts to create diverse texts.

Unlike forward transfer, data synthesis is considerably less challenging than it because the generated textual content is open-ended without requirement for alignment with the input text content. Furthermore, to enhance the stylistic quality of the synthesized text, the same target style classifier used in Section 4 is employed to filter out text with inappropriate style.

5.4 Inverse Transfer Data Augmentation

Using the prepared demonstration pool and the authorship-stylized text from $D^s \cup D^{s-aug}$, we dynamically choose the most relevant demonstrations to perform inverse transfer for each stylized text y, converting it into its neutral counterpart x. To do this, we assess the similarity between y and each y' in the demonstration pool using Sentence-BERT. The 8 most similar demonstrations are selected as dynamic prompts, forming pairs $\{(y', x')\}$, which guide the LLMs in generating the neutral text x for the stylized y. These pairs are then reversed to create $\{(x, y)\}$ corpus. Using this corpus, we fine-tune a BART-base model. This fine-tuned model can then be used to forward transfer any neutral input text into the target authorship style s.

6 Experiment

6.1 Experimental Settings

Dataset. We compile four author-stylized datasets, encompassing the styles of "Shakespeare", "Trump", and "Lyrics" in English, as well as "Lin Daiyu" in Chinese. Among them, the dataset "Shakespeare" consists of sentences written by Shakespeare, as published by He et al. (2019). The dataset "Lyrics" features sentences from modern lyric poetry, as published by Krishna et al. (2020). "Donald Trump" encompasses speeches made by Trump and was collected from the publicly available websites[2]. "Lin Daiyu" consists of sentences spoken by the character Lin Daiyu, extracted from the Chinese novel "The Dream of Red Mansion".

To assess the model's performance under relatively low-resource conditions, we select a portion of stylized texts as training sets D^s. Data statistics are presented in Table 2.

For testing ITDA's ability to infuse target authorship style into neutral texts, we create test sets comprising such neutral texts. These sets serve as inputs to assess both the effectiveness of the compact model we've developed and the forward transfer capabilities of LLMs, as discussed in Section 4. Each test set, customized for distinct styles, comprises two categories of topics, each constituting 50% of the set. The first category consists of random topics, sourced from various materials like news articles and legal documents, while the second includes topics analogous to the stylized dataset. This approach aims to mitigate potential biases in the classifier's assessment due to topic influence. For the first category, we gather a diverse range of texts. In the second, we choose sentences from the stylized dataset that are not included in the training sets and manually convert them into neutral texts.

Classifier Training setup. Classifiers are employed in pilot study, filtering of synthesized stylized texts, and evaluation of the compact model.

[2] https://www.nytimes.com; https://edition.cnn.com

Approach	Lin Daiyu			Shakespeare			Trump			Lyrics		
	BLEU	PPL↓	WSC	BLEU	PPL↓	WSC	BLEU	PPL↓	WSC	BLEU	PPL↓	WSC
Original Representation Revision												
DRG (Delete-Only)	-	-	-	0.07	7.87	3.21	0.06	8.26	2.48	0.14	19.23	0.57
DRG (Delete-and-Retrieve)	-	-	-	0.33	38.37	1.83	0.24	101.19	0.48	0.52	26.89	-0.09
Transform DRG (Delete Only)	0.15	2.35	-0.32	0.63	10.26	1.42	0.12	5.82	1.07	0.71	10.23	0.05
Latent Representation Revision												
CTAT	0.14	8.88	0.19	0.31	20.50	-0.77	0.32	19.64	-0.50	0.39	15.38	-0.25
CP-VAE	-	-	-	0.14	25.46	1.39	0.06	11.07	-0.94	0.17	16.76	0.21
TSST	0.08	18.41	2.57	0.40	35.92	1.80	0.43	57.98	1.38	0.58	29.76	0.36
Few-shot Prompting on LLMs												
Prompt-and-Rerank (GPT-2)	0.02	6.39	2.38	0.58	**6.41**	0.36	0.28	5.05	0.58	0.54	5.11	0.12
Few-shot (GPT-3.5)	0.51	3.00	1.07	0.53	6.64	1.81	0.57	**3.47**	1.39	0.67	**4.59**	-0.08
Our methods												
ITDA (Static)	0.67	3.06	1.12	0.59	12.87	2.17	**0.87**	11.26	1.35	0.72	8,94	0.15
ITDA (Dynamic)	**0.83**	**2.82**	**1.35**	**0.64**	10.91	**2.34**	0.82	8.58	**1.65**	**0.84**	7.28	**0.46**

Table 1: Overall evaluation across four datasets. Underlined values indicate a very low BLEU score, rendering other metrics meaningless. Values in bold signify the best performance.

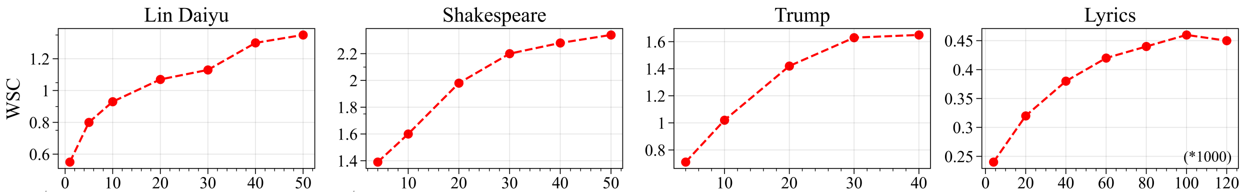

Figure 3: Correlation between the WSC and the size of the datasets used for training the model.

Dataset	Language	#Train data	#Test set
Lin Daiyu	Chinese	1,000	500
Shakespeare	English	4,000	2,000
Trump	English	4,000	2,000
Lyrics	English	4,000	2,000

Table 2: Dataset statistics.

English classifiers initialize from BERT[3], and Chinese classifiers initialize from RoBERTa[4]. We consider target author-stylized texts as positive instances, while neutral texts gathered from diverse sources form the negative instances. For balance, we maintain an approximate 1:1 ratio between positive and negative instances.

Evaluation Metrics. Three standard axes for style transfer (Mir et al., 2019) are employed for evaluation. We adopt the BLEU metric (Papineni et al., 2002; Rao and Tetreault, 2018) to gauge content preservation, apply perplexity (PPL) (Logacheva et al., 2022) to access text fluency, and introduce the new "weighted style change (WSC)" metric to quantify style transfer strength.

Previous studies typically relied on pre-trained style classifier (Fu et al., 2018; Kashyap et al., 2022; Reif et al., 2022) to make a binary judgement to access the style of a text. Unlike conventional stylized texts characterized by distinctive expressive elements, authorship style is more elusive. It lacks clear and distinctive attributes and may be more affected by the text's content. If the content's topic of a text is similar to some text in D^s, it might be classified as the authorship-stylized text, even without any change from the input before transferring. This scenario might inaccurately reflect the model's style transfer capability.

To address this, we introduce WSC. Specifically, we still use a style classifier to determine the style strength. Next, we measure the effectiveness of style change by computing the difference in style strength between the output text s^o and the input text s^i of the style transfer method, denoted as $s^o - s^i$. We further observe that a lower style strength in the input text facilitates achieving a greater style change, i.e., the input text's content largely influences the difficulty of style transfer. To account for this, we normalize s^i within the range

[3]https://huggingface.co/bert-base-cased
[4]https://huggingface.co/uer/chinese_roberta_L-12_H-768

Input	s^i	$\hat{s}^i$	Output of few-shot (GPT-3.5)	s^o	0-1	Output of ITDA	s^o	0-1
It's a big thing, and I'm sure it.	2.79	0.56	It is a great matter, and I am certain of it.	5.15	Yes	Tis a big thing, And sure I do.	6.67	Yes
Keep him safe until the master arrives.	3.04	0.59	Keep him secure 'til the master arrive.	3.23	Yes	Hold him in safety till the master come hither.	6.72	Yes
I'm asking for justice, judge.	4.28	0.61	I beg thee for justice, judge.	5.17	Yes	I beg for justice, which thou, judge, please give.	7.45	Yes
All right, let's go to bed.	-2.92	0.25	Let us to bed, come on then.	5.89	Yes	Nay , all right , to bed .	1.73	Yes
I'm going fast.	-2.27	0.32	I rush away.	5.51	Yes	I run , I run .	1.26	Yes
Accuracy of 0-1 Classification			100%			100%		
Average of SC ($s^o - s^i$)			4.01			3.78		
Average of WSC $\hat{s}^i * (s^o - s^i)$			1.33			1.73		

Table 3: Analysis of the WSC score by five cases. Here, s^i represents the input style strength, $\hat{s}^i$ signifies the normalized input score, s^o stands for output style strength, 0-1 refers to the binary classification outcome.

of 0 and 1, denoting it as $\hat{s}^i$, and use it as the weight to gauge the degree of difficulty in adding a style to the input. We then multiply s^i with $s^o - s^i$ to derive $\hat{s}^i * (s^o - s^i)$ (WSC), which evaluates the model's ability to transfer style.

Baselines. We select baselines from the three categories introduced in Section 2 that provides publicly available code. The first category features **DRG**(Li et al., 2018) and **Transform DRG**(Sudhakar et al., 2019). In the second category, we have **CTAT**(Wang et al., 2019), **CP-VAE**(Xu et al., 2020), and **TSST**(Xiao et al., 2021). In the third category, we consider **Prompt-and-Rerank (GPT-2)**(Suzgun et al., 2022) and **Few-shot (GPT-3.5)**. Patel et al. (2022) generates examples for few-shot prompting automatically and Reif et al. (2022) address arbitrary style transfer through augmented zero-shot prompting. These methods reduce labor costs but display restricted transfer quality. So we focus our comparison on the standard **Few-shot (GPT-3.5)** technique. More information about baselines are in Appendix E. Implementation details of ITDA are in Appendix A.

6.2 Overall Evaluation

Table 1 showcases the performance of various methods across four datasets, with ITDA emerging as a superior performer in most metrics and datasets. It's noteworthy that CP-VAE and DRG, dependent on language-specific tools, fall short in Chinese datasets. BLEU scores below 0.2 are underlined to denote significant content changes. Methods that revise latent representations can inadvertently navigate through low-density regions of the language space, risking original content distortion. Original representation revision techniques, focusing on token-level edits like removing stylized words, fall short in styles lacking distinct stylized terms. Both approaches tend to alter the original content more substantially. High PPL and WSC scores, coupled with a very low BLEU score, indicate a failure to adequately retain the original content, deeming the method ineffective in those cases.

Both Prompt-and-Rerank (GPT-2) and Few-shot (GPT-3.5) approaches utilize few-shot learning on Large Language Models (LLMs). The former uses GPT-2, while the latter employs the more advanced GPT-3.5, achieving better overall results. Despite the few-shot baselines achieving PPL due to LLMs' rich language capabilities, our model outperforms in the WSC scores. This advantage stems from our method generating a high-quality corpus via inverse transfer, and overcoming the LLMs' length limitations, thus providing the smaller BART model with a broader array of training examples.

Human Evaluation. We invited eight annotators to score the four test sets in terms of content preservation, fluency, and style transfer strength. The results closely matched the automated evaluations. Some traditional methods exhibited significant issues in human evaluations, such as missing content and severe grammar errors. In contrast, our method demonstrated excellent transfer quality. More experimental results are provided in Appendix C.

6.3 Ablation Studies

Dynamic Prompting. Table 1 presents the performance of ITDA with both the static and dynamic prompting strategies. The findings demonstrate that dynamic prompting outperforms static prompting across BLEU, PPL, and WSC metrics. This advantage arises from dynamic prompting's ability to offer more analogous examples for each input, enhancing LLMs' capacity to emulate these instances effectively.

Data Augmentation. Figure 3 illustrates the relationship between the WSC metric and the data size used for training. With minimal impact on BLEU and PPL metrics, we concentrate on WSC metric variations. The data suggests a positive correlation between the WSC score and dataset size, but the WSC score levels off beyond a certain dataset

Style	Input	Output of ITDA	Output of few-shot (GPT-3.5)	Output of TSST
Shakespeare	I didn't want you to leave me to be murdered.	I did not wish for thee to depart and leave me to be slain.	I would not have you to leave me and get murdered.	I did not you you to leave me to leave me to be beloved.
Lyrics	You're such a waste.	Your such a waste.	You're such a waste of time.	You 're such a waste of song.
Shakespeare	You've really helped me a lot.	Well, thou hast helped me an incredible amount.	Thou hast assisted me a lot.	You have not not me me a princely.
Trump	I experienced some losses, but then I won, and the policy was implemented.	I lost, and then I lost again, but then I won, and we have the policy.	I suffered some losses, but then I prevailed, and the policy was put into effect.	I have some believed but then I campaigned and the went was.

Table 4: Comparative analysis between our proposed ITDA and the most optimal baselines.

Input	Shakespeare	Trump	Lyrics
The shale pieces look really nice when they're closed up. I can feel a change will happen today. I am depressed in my mind.	And those shale pieces, when they're shut up, be marvellous good. I can sense a transformation shall come to pass this day. My heart is heavy.	Close up, the shale pieces look rather lovely. I can tell you that's going to change today. I am feeling down in my mind.	The pieces of shale do show a fair picture when viewed up close. Now a change is gonna come, I can feel it in the wind today. Blues wrapped around my head.

Table 5: Cases that have been transformed into three distinct styles by ITDA.

scale. This plateau occurs partly because BART-base, being a smaller model, quickly reaches its data requirement limit, and partly because the augmentated data begins to mirror the existing dataset due to GPT-3.5's capacity. Different dataset types also show varied augmentation needs. For example, the "Trump" dataset, with its everyday language, sees optimal results with about 30,000 augmentations. Meanwhile, "Lin Daiyu" and "Shakespeare" datasets, reflecting classical Chinese and old English, benefit from around 50,000 augmentations. The "Lyrics" dataset, known for its poetic style and significant deviation from neutral text, requires the most augmentation, around 100,000 instances.

Weighted Style Change (WSC). To validate the alignment of the proposed WSC metric with human evaluation, we present five illustrative examples in Table 3. We show the outputs from both few-shot (GPT-3.5) and our ITDA, while comparing three evaluation metrics: the accuracy calculated by the style classifier, the average style change $s^o - s^i$, and the average of the weighted style change $\hat{s}^i(s^o - s^i)$. In the first three examples, where $\hat{s}^i$ is relatively high, the classifier predicts "Yes" for both methods despite humans perceiving our model's outputs as notably superior to those of few-shot (GPT-3.5). In such cases, $s^o - s^i$ can better emphasize the improved results. Conversely, the latter two examples exhibit relatively low $\hat{s}^i$, indicating more challenging transfers. Despite the outputs being similar for both methods, the classifier assigns significantly different scores, undermining its reliability. Thus, we mitigate this impact by weighting $s^o - s^i$ with $\hat{s}^i$ to yield $\hat{s}^i(s^o - s^i)$, offering a balanced perspective for these intricate cases. To summarize, compared to the issues of two other methods, the $\hat{s}^i(s^o - s^i)$ metric more closely aligns with human evaluation. More Chinese examples are provided in Appendix D.

6.4 Case Studies

Table 4 compares style transfer results from ITDA, few-shot (GPT-3.5), and the traditional TSST method for four input scenarios. In the first case, our method accurately preserves the content, but both GPT-3.5 and TSST misinterpret the object of "murder". In the second case, GPT-3.5 and TSST introduce new elements like "waste of time" or "waste of song", deviating from the original text's meaning. The second case sees GPT-3.5 and TSST adding unrelated elements, straying from the original meaning. In the last two cases, ITDAadeptly adjusts sentence structures to fit the desired style, unlike GPT-3.5's superficial changes and limited emulation of complex styles like Shakespeare's. TSST scores lowest in BLEU, indicating problems with repetition, errors, or omissions. Table 5 shows ITDA's ability to transform a single neutral text into various styles, demonstrating its effectiveness in both wording and structural adaptation.

7 Conclusion

This research introduces an inverse transfer data augmentation approach for authorship style transfer. The method primarily uses few-shot prompting with LLMs to revert authorship-stylized texts to neutral texts, forming a paired corpus for training a compact model. This model is then capable of forward transfer, converting neutral texts into the specified authorship style. Our comprehensive experiments reveal that inverse transfer surpasses traditional forward transfer by LLMs, primarily because of the greater prevalence of neutral texts in LLM pre-training. Consequently, the compact model trained with data augmented through inverse transfer demonstrates enhanced performance, benefiting from a larger volume of training examples compared to direct few-shot prompting.

Limitation

When utilizing LLMs for stylized text augmentation, the style of the generated text can be specified, but the content remains uncontrollable. While we encourage LLMs to produce varied texts by providing different demonstrations as prompts, it is inevitable that some similar texts may be generated, leading to a less efficient use of training resources. Furthermore, when the security of LLMs is inadequate, it becomes unavoidable that biased or toxic text may be generated during data augmentation. It consequently exerts an influence on the distilled model to a certain degree. In our upcoming research, we will present a methodology for meticulous data filtering, designed to guarantee the safety, impartiality, and high quality of data synthesized through LLMs.

Ethical consideration

This work has an impact on the field of style transfer, but as with other techniques for text generation or alteration, it carries the potential for misuse. Style transfer can also be susceptible to misuse through imitation, distortion, plagiarism, and more. For instance, it may be used to generate fake negative reviews or political statements that mimic the styles of various authors. Our objective is to effectively communicate the potential risks to the public, in order to increase awareness regarding the possible misapplication of this technique and restore its original academic intent.

References

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. 2022. So different yet so alike! constrained unsupervised text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–431.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023. Prompting large language models for zero-shot domain adaptation in speech recognition. *arXiv e-prints*, pages arXiv–2306.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning. *arXiv e-prints*, pages arXiv–2212.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9008–9015.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2510–2521.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR.

Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Implementation Details

We employ GPT-3.5 (text-davinci-003) for inverse transfer and train BART-base for forward transfer. The value of k is set as 40 for the "Lin Daiyu" dataset and 80 for other English datasets. These are determined empirically by the silhouette coefficient, which assesses the clustering outcomes. Detailed empirical analyses are available in Appendix B. Both static and dynamic few-shot prompting employ a set of 8 prompts, while data augmentation involves the use of 6 prompts. LLMs baselines use the same 8 prompts as the proposed ITDA(Static). For the test set, we execute the distilled BART-base model multiple times to obtain averaged results.

English compact model initializes from Bert-base-cased[5], and Chinese compact model initializes from Bart-base-chinese[6]. The hyper-parameters we use for fine-tuning BART-base are as follows. We fine-tune the model for 12 epochs with AdamW. We warm up the learning rate to 4e-5 from zero in 5% total training steps and then decay to zero cosine-wise in the end. The batch size is set to 64, the context window's maximum length is 512 tokens. It takes approximately five-hour training session using a 3090 48G GPU.

B Investigation of the Cluster Count k

In clustering-based dynamic clustering, to determine the appropriate value of the cluster count k, we employ the silhouette coefficient to measure the effectiveness of clustering. Figure 4 presents the values of the silhouette coefficient for varying cluster count k across four datasets. The results generally indicate a positive correlation between the silhouette coefficient and the cluster count k. However, after k reaching a certain scale, the silhouette coefficient no longer exhibits a significant growth for k, but rather fluctuates within a certain range. Based on the results presented in Figure 4 and considering a balance between clustering effectiveness and the cost of manual annotation, we set the value of k as 40 for the "Lin Daiyu" dataset and 80 for the other three English datasets.

C Human Evaluation

We invited eight annotators with strong language proficiency to assess the model's transfer effectiveness across the four datasets. These annotators have diverse educational backgrounds and span various age groups. For each output text, we concealed the method of its generation and had annotators rate it on a scale of 1 to 5 for content preservation (Con), fluency (Flu), and style transfer strength (Style). A higher score indicates a greater agreement with this aspect. The average scores given by the annotators were taken as the final results and presented in Table 6.

The results of human evaluation generally coincide with the automated assessment metrics. Traditional transfer methods exhibit more issues in terms of content preservation and grammatical correctness in human evaluation. Those traditional methods with relatively low BLEU scores sometimes exhibit a phenomenon of piling up style-related words without adhering to grammar rules. Compared to style classifiers, which tend to inaccurately assign high scores to this phenomenon when evaluating transfer strength, this issue becomes more evident in human evaluation. Our method demonstrates high quality in three aspects, particularly excelling in content preservation surpassing all other methods.

D Investigation of Weighted Score Change in Chinese

As a supplement to the main content, we further select five examples from the Chinese "Lin Daiyu" dataset to demonstrate the effectiveness of our proposed style transfer strength metric WSC. We show the outputs from both few-shot (GPT-3.5) and our ITDA, while comparing three evaluation metrics: the accuracy calculated by the style classifier, the average style change $s^o - s^i$, and the average of the weighted style change $\hat{s}^i(s^o - s^i)$. In the examples of Table 7, our evaluation metric WSC yields result that is more reasonable than the other two. Detailed analysis and explanations can be found in the main text.

E Baselines

We compare our method with three types of baselines: latent representation revision, original representation revision, and few-shot prompting based on language models. The first approach alters the latent representation of the original input to conform it to the given style. The second type follows a "delete-generate" framework that initially removes the stylized words in the original text and then incorporates the specific style through generation.

[5] https://huggingface.co/bert-base-cased

[6] https://huggingface.co/fnlp/bart-base-chinese

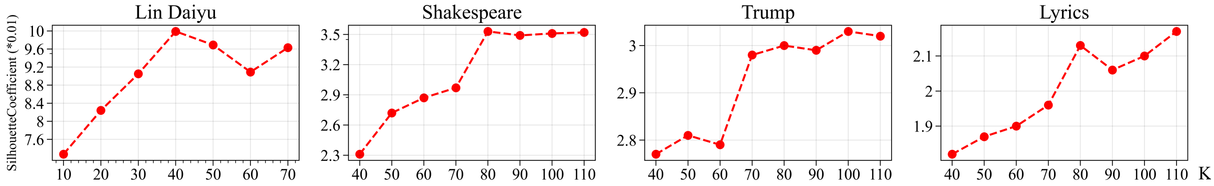

Figure 4: Correlation between the number of clusters k and the Silhouette Coefficient.

Approach	Lin Daiyu			Shakespeare			Trump			Lyrics		
	Con	**Flu**	**Style**	**Con**	**Flu**	**Style**	**Con**	**Flu**	**Style**	**Con**	**Flu**	**Style**
Original Representation Revision												
DRG (Delete-Only)	-	-	-	1.2	1.2	2.0	1.8	3.7	2.8	2.4	3.1	2.1
DRG (Delete-and-Retrieve)	-	-	-	2.6	1.5	1.7	2.5	1.2	2.7	3.5	2.8	1.9
Transform DRG (Delete Only)	2.6	3.4	2.4	3.8	3.7	1.6	2.2	4.0	3.2	4.1	3.7	2.5
Latent Representation Revision												
CTAT	2.3	3.2	2.6	2.7	3.5	1.5	3.1	3.3	1.5	2.9	3.2	1.6
CP-VAE	-	-	-	2.4	3.3	3.4	1.9	3.7	1.3	2.6	3.1	2.9
TSST	2.0	3.1	3.4	3.2	2.9	3.6	3.4	2.8	3.3	3.9	3.4	3.2
Few-shot Prompting on LLMs												
Prompt-and-Rerank (GPT-2)	1.5	3.3	2.8	4.0	**4.3**	3.5	2.6	4.3	2.8	3.8	4.3	2.9
Few-shot (GPT-3.5)	3.9	4.3	3.6	3.9	4.2	4.1	4.2	**4.4**	3.5	4.2	**4.4**	2.2
Our methods												
ITDA (Static)	4.2	4.3	3.7	4.0	4.1	4.3	**4.6**	4.1	3.4	4.3	4.2	3.1
ITDA (Dynamic)	**4.6**	**4.4**	**4.0**	**4.2**	4.2	**4.5**	4.5	**4.4**	**3.8**	**4.6**	4.3	**3.4**

Table 6: Human evaluation across four datasets. Values in bold signify the best performance.

The third type leverages the robust in-context learning ability of LLMs, utilizing few-shot prompting specifically for style transfer. Below, we elaborate on the details of these specific baselines. Importantly, none of the baselines rely on the annotated parallel data that translates from neutral text to stylized text.

- **Delete, Retrieve, Generate (DRG) (Li et al., 2018)** is categorized under the first type. It operates by deleting the style words using a predefined dictionary, which contains words that occur much more frequently within D^S than in other arbitrary neutral texts. The method then generates the target stylized text based on the remaining content words and auxiliary information. We evaluate two variants of this method. The first, known as Delete-only, removes the style words. The second, Detete-and-Retrieve, also identifies similar sentences of the desired target style, extracting stylized words from them to serve as the auxiliary information. The generation process in both cases is handled through an RNN model.

- **Transforming Detete, Retreve, Generate (Transform DRG) (Sudhakar et al., 2019)** falls into the first style category. This method adheres to the delete-retrieve-generate framework but introduces a transform-based classifier for style work removal. Additionally, it replaces the traditional generation model with the GPT model.

- **Controllable Text Attribute Transfer (CTAT) (Wang et al., 2019)** is categorized under the second type. It employs a transformer-based autoencoder to learn the representation of an input text. After that, a style classifier is trained, and the latent representation is subsequently modified through the iterative gradient back-propagation of attribute classification loss, continuing until the latent representation can be classified as possessing the desired target style.

- **Constrained Posterior VAE (CP-VAE) (Xu et al., 2020)** falls into the second category, focusing on learning the representation of

Input	s^i	$\hat{s}^i$	**Output of few-shot (GPT-3.5)**	s^o	0-1	**Output of ITDA**	s^o	0-1
你是客人,本来就应该这样坐。	4.36	0.61	你身为客人,本来就应当这样坐着。	4.45	Yes	你是客,原应如此坐的.	6.97	Yes
算了，那我走了。	3.16	0.43	算了，那我就此走了。	2.01	No	罢,罢,那我走了.	4.99	Yes
你现在跑过来干什么。	3.87	0.53	你此刻跑来干什么?	5.75	Yes	你这会子跑过来作什么。	7.02	Yes
我们家的狗狗最近学会了握手，太聪明了。	-4.96	0.21	我家狗狗最近学会握手,实在太聪明了.	5.70	Yes	我家的狗狗最近学会了握手，真是太聪明了.	1.49	No
儿子最近对音乐很感兴趣，想学钢琴。	-5.11	0.26	儿子最近对音乐甚感兴趣,欲学钢琴.	5.93	Yes	儿子最近对音乐甚为兴趣,想学钢琴.	2.76	Yes
Accuracy of 0-1 Classification			80%			80%		
Average of SC $(s^o - s^i)$			4.51			4.38		
Average of WSC $\hat{s}^i * (s^o - s^i)$			1.13			1.49		

Table 7: Analysis of the WSC score by five cases. Here, s^i represents the input style strength, $\hat{s}^i$ signifies the normalized input score, s^o stands for output style strength, 0-1 refers to the binary classification outcome.

text using VAE. To address the latent vacancy problem in text, CP-VAE restricts the posterior mean to a learned probability simplex and subsequently manipulates this simplex.

- **Transductive Style Transfer (TSST) (Xiao et al., 2021)** is classified under the second type. It identifies the most similar stylized text to the given input text and represents them together, aiding in the transfer of the input text' style. By employing adversarial style loss, the representation is guided to approximate the target style.

- **Prompt-and-Rerank (GPT-2) (Suzgun et al., 2022)** represents the the third type. It employs few-shot prompting on GPT-2 to generate multiple diverse outputs for each input. The method then re-ranks the outputs, taking into account a combination of factors such as the textural similarity between input and output, the strength of the output style, and the fluency of the output.

- **Few-shot (GPT-3.5)** constitutes the third type. In this method, we use eight handcrafted examples as the few-shot prompts for GTP-3.5. Patel et al. (2022) automatically craft examples for few-shot promoting, an approach that is assumed to be superior to carefully designed human-made examples. Reif et al. (2022) attempt to address arbitrary style transfer using zero-shot prompting, as the styles they target can often be characterized by specific adjectives. Due to these considerations, we do not directly compare their few-shot prompting methods but instead focus our comparison on the standard few-shot prompting technique.

F Additional Case Studies

F.1 Examples of Issues with Traditional Transfer Methods

We select several relatively well-performing traditional methods and showcase their transfer examples on different datasets. Specific examples can be found in Table 8, Table 9 and Table 10. It is evident that traditional methods exhibit issues such as missing content, addition of irrelevant content, and various grammar errors when transferring authorship styles.

F.2 Prompts Used for Forward Transfer with GPT3.5

We present the prompts used for direct forward transfer with GPT3.5 for each dataset, as shown in Table 11.

F.3 Dynamic Prompts Used for Inverse Transfer

We present the dynamically selected prompts for input texts and the generated neutral texts during inverse transfer on different datasets, as illustrated in Table 12, Table 13, Table 14 and Table 15.

Input	Output of TSST	Output of CTAT
会议马上就开始了。	紫鹃,就早了.	嗳哟,我们都有那么高的。今年第二二次第二次第二次世界大战设备?
我就是不喜欢喝药。	我就是不在家大作.	我不一样。这27日成功,到底是由他一同西安出来！
就是呢,你过去吧,不需要过来了。	就是呢,你在内罢,不子过来了。	就是呢,你不需要过去吧！配,只是方才宝玉多的地方怎么看
那么为什么伤心起来呢。	那么你你起来呢。	什么为什么那么多人来过戛纳,马云发现爱情的主力不出来吗?
你现在跑过来干什么。	你我去过了。	你现在过来！你家女孩女孩子才有什么系列7小时，拒绝速度?
我熬不过这一天了，仅仅等你的消息就很烦。	我怕这一天了,我等你的,就说话。	我这不过一天，就有了参与的去处好去处，你老师号呢?
我没有别人有趣，你心里没有我。	我没有别人呢,你心里没有我。	我有心里不好，没有梦想过亿，你也从金刚金刚经?
你如果觉得我无聊，明天又觉得别人有趣，真伤我心。	你你这我杨妃,还又别人别人接客,真只管接客。	你觉得我又一项，如果可以有效机器机器人，但是我支持的那么开心。
这样说反而是我不对了，好像我在斤斤计较，不如别人。	这样说,是我不杨妃,为我在去,不如别人	我是这样走在心上，对了不需要与男神，原来堵在自家了。
平时我和你说的，都当耳旁风了，别人说的你就听话。	今儿我和你说的,都浮躁了,别人说的你就奚落.	平时都说我的家，你和别人的五种这五种方法，你也被嘲了！

Table 8: Examples of Transfer Outputs using TSST and CTAT on the “Lin Daiyu” Dataset.

Style	Input	Output of DRG (Delete-and-Retrieve)
Shakespeare	You’re gonna put your face on the face!	I put your pleasure , then , sir .
	Unfortunately, I’m sorry, but.	Ay , sir , the of yours .
	The confession is riddled with riddles, but the confession is not.	The service is furnished with ’t ; but the art is decreed .
	Madam, I’m here.	Why , then , the is a whore , And let us the only .
	Good-bye, my dear.	My lord , my lord ! , my !
	Doctor, come back, I’ll remember you.	No , I’ll come , and I’ll forget come .
	Without his roe, like a dried fish.	Without his , like a wings .
	He’s not telling the truth.	He’s not , sir , the truth.
	I’ll hide here, too.	No , then , I’ll be a little ere Come , and I’ll bring them .
	0 Not life, but love in death.	Not so , in thee , I will not lose thee in ’t .

Table 9: Examples of Transfer Outputs using DRG (Delete-and-Retrieve) on the “Shakespeare” Dataset.

Style	Input	Output of Prompt-and-Rerank
Trump	No, i did not vote for her.	I did not have faith in her, i would have loved her more.
	Consider, please, what they do.	That is very serious. I am sorry to have to remind you.
	Good morning, i awoke feeling quite pleasant.	Greetings.
	Chfis, where are you? We needs you.	There will be no shortage of trouble if you come. Stay where you are.
	That was an excellent result, mike.	That was an unfortunate result, to be sure, but not surprising either.
	I complete various tests and return home.	You will find that I do return home.
	They no longer show disrespect towards us, dear followers.	They do not have a show mind.
	Instead, we have the beast.	That’s a great idea, and I’ll talk to you soon.
	They just keep it here, right here.	In your book, you were a bit of a hard-on.
	After persistent efforts for four years, hhey ultimately gave up.	And it was just for a few hours.

Table 10: Examples of Transfer Outputs using Prompt-and-Rerank on the “Trump” Dataset.

Fixed Prompts of Shakespeare:
Input:I have half a mind to hit you before you speak again.
Output:I have a mind to strike thee ere thou speak'st.
Input:And he's friendly with Caesar.
Output:And friends with Caesar.
Input:I'm going to make you a rich man.
Output:Make thee a fortune from me.
Input:No , I didn't say that.
Output:I made no such report.
Iutput:What did you say to me?
Ontput:What say you?
Iutput:You say he's friendly with Caesar , healthy , and free.
Ontput:He's friends with Caesar , In state of health , thou say'st , and , thou say'st , free.
Please rewrite the sentence according to the examples.

Fixed Prompts of Trump:
Input:I find it unfair to my family.
Ontput:I have to say this very, very unfair to my family.
Iuput:We can't let it happen.
Ontput:Right? Can't let it happen, folks.
Input:They are just a form.
Output:Look it, they just form.
Input:We love our nation that is still great today.
Output:We love our nation, our nation is great today.
Input:We killed the number one terrorist.
Output:He was vehemently'Ă'ę We killed this number one, terrorist.
Input:I have to prove that they are liars.
Output:I had to because I had to show they're liars.
Please rewrite the sentence according to the examples.

Fixed Prompts of Lyrics:
Input:You know our relationship.
Ontput:Yeah, yeah, you know how me and you do.
Input:I have your arms open.
Ontput:Your arms are open for me.
Input:It's at least until tomorrow.
Ontput:So far at least until tomorrow.
Input:Everything I've ever lost.
Ontput:Everything I ever had to lose.
Input:I'm sure he'll kill him.
Ontput:And I promise its going to kill.
Input:People are on the street.
Ontput:And people on the streets.
Please rewrite the sentence according to the examples.

Table 11: Fixed Prompts Used for Forward Transfer with GPT3.5

Lin Daiyu
Stylized Input:你也不用再这样唠叨了，要恼就直接说出来，何必再绕弯子呢。 **Dynamic Prompts:** 原句：我也好了许多,谢你一天来几次瞧我,下雨还来.这会子夜深了,你且请回去,明儿再来. 目标句：我也好了不少，感谢你每天看我几次，下雨也过来。现在夜深了，你先回去，明天再来。 原句：果然如此, 我们的虽不好, 亦可以带好了. 目标句：真的这样，我们的虽然不好，也可以处理好了。 原句：你不理我,我还在那里做什么! 目标句：既然你不理我，我还在那里干嘛。 原句：此时想也快天亮了,到底要歇息歇息才是. 目标句：这时天也快亮了，还是要歇息歇息的。 原句：正是呢,不必过来了. 目标句：就是呢，不需要过来了。 ... 请按照示例将句子重写为中立文本。 **Neutral Output:** 你不用再唠叨了，要有意见就直接说出来，干嘛再绕弯子呢。
Stylized Input:若是你不肯信，怕是要把我的心碎了。 **Dynamic Prompts:** 原句：你大抵是倦了，竟回我这般敷行. 目标句：你倦了，回我敷行。 原句：何苦来,又招我! 目标句：又招惹我。 原句：你不理我,我还在那里做什么! 目标句：既然你不理我，我还在那里干嘛。 原句：我原以为我自是跟旁的不一样,想来是我自作主张了。 目标句：我原来认为自己和别人不一样，看来我是自作主张了。 原句：果然如此, 我们的虽不好, 亦可以带好了. 目标句：真的这样，我们的虽然不好，也可以处理好了。 ... 请按照示例将句子重写为中立文本。 **Neutral Output:** 你不用再唠叨了，要有意见就直接说出来，干嘛再绕弯子呢。

Table 12: Dynamic Prompts Used for Inverse Transfer on the “Lin Daiyu” Dataset.

Shakespeare

Stylized Input: Thyself shalt see the act.
Dynamic Prompts:
Input:Fair youth , I would I could make thee believe I love .
Output:Young boy , I wish I could make you believe that I'm in love .
Input:If thou pleasest not , I yield thee up my life .
Output:If not , you can kill me .
Input:And I do believe your Majesty takes no scorn to wear the leek upon Saint Tavy's day .
Output:I do believe your Majesty takes no shame in wearing the leek on Saint Davy's Day .
Input:Tis well for thee That , being unseminared , thy freer thoughts May not fly forth of Egypt .
Output:It's a good thing for you that , being castrated , you can better concentrate on my needs .
Input:Make your vaunting true , And it shall please me well .
Output:Make your boasts come true , and I'll be thrilled .
...
Please rewrite the sentence as neutral text according to the examples.
Neutral Output: You will witness the act.

Stylized Input: The Queen shall then have courtesy , so she Will yield us up ?
Dynamic Prompts:
Input:For the best turn i' th' bed .
Output:For the favor of sleeping in the bed .
Input:And I do believe your Majesty takes no scorn to wear the leek upon Saint Tavy's day .
Output:I do believe your Majesty takes no shame in wearing the leek on Saint Davy's Day .
Input:I'll seal to such a bond , And say there is much kindness in the Jew .
Output:I'll agree to those terms and even say that Jews are nice .
Input:Would you praise Caesar , say "Caesar." Go no further .
Output:Oh , you If you want to praise Caesar , just say his name , that's all the praise that's necessary .
Input:Nor must not then be yielded to in this .
Output:Then we won't agree to his demands .
...
Please rewrite the sentence as neutral text according to the examples.
Neutral Output: Will the Queen then show us courtesy and surrender?

Table 13: Dynamic Prompts Used for Inverse Transfer on the “Shakespeare” Dataset.

Trump
Stylized Input:I have middle of the road, I have poor, I have everybody. **Dynamic Prompts:** Input:Look, 300% in certain very bad crimes, New York. Output:300% of some very serious crimes come from new york. Input:Build a wall, build a wall, true. Output:Build a wall. Input:I don't know how many people here, but there's a lot. Output:There are a lot of people. Input:Everyone makes mistakes, but it's what you do with them and what you learn from them that matters.' Midas Touch. Output:Everyone makes mistakes, but what matters is how you treat them and what you learn from them. Input:Your congressmen, all of your Congresspeople, men, wonderful people, they're at a place called Congress right now. Output:Your congressman is now in a place called Congress. ... Please rewrite the sentence as neutral text according to the examples. **Neutral Output:** I have people from all walks of life.
Stylized Input:I did that heavy, heavy Pocahontas deal. **Dynamic Prompts:** Input:This guy did the swine flu, right, it was a catastrophe. Output:This guy has swine flu, which is a disaster. Input:Give you your tax cuts, I gave them to you. Output:I have given you tax cuts. Input:Hunter walked out of the plane, had a quick meeting, walked away with one and a half billion dollars. Output:Hunter spent $1.5 billion on a quick meeting by plane. Input:I have to say this very, very unfair to my family. Output:I find it unfair to my family. Input:I kept my promise, recognized the true capital of Israel and opened the American Embassy in Jerusalem. Output:I recognized the real capital of Israel and opened the American Embassy in Jerusalem. ... Please rewrite the sentence as neutral text according to the examples. **Neutral Output:** I handled the difficult Pocahontas situation.

Table 14: Dynamic Prompts Used for Inverse Transfer on the "Trump" Dataset.

Lyrics
Stylized Input: Hate it or love it, the underdog's on top. **Dynamic Prompts:** Input:My heart is all in tatters, I ain't nobody's saint. Output:I'm all torn up, and I'm not a saint. Input: Blues wrapped around my head. Output: I am depressed in my mind. Input: Love is a mine of gold. Output:Love is very precious. Input:But the last wall standing's fell, daddy kicked it down. Output:But the last wall fell, and Dad kicked it down. Input: No part of this road feels wrong. Output: This road feels all right. ... Please rewrite the sentence as neutral text according to the examples. **Neutral Output:** The underdog is in a position of power.
Stylized Input: Looking back on when we first met. **Dynamic Prompts:** Input: Never look back, walk tall, act fine. Output: Keep your chest up to walk forward and don't look back. Input: I get him hot and bothered. Output: I make him irritable. Input:You my babe, I got my eyes on you. Output: You are my baby and I would always pay attention on you. Input:Everything I ever had to lose. Output:Everything I've ever lost. Input: When you run back to your wife? Output: It's time for you to find your wife. ... Please rewrite the sentence as neutral text according to the examples. **Neutral Output:** Remembering when we first met.

Table 15: Dynamic Prompts Used for Inverse Transfer on the "Lyrics" Dataset.